UNA PROPUESTA Y UN ETIQUETADOR DE CODIFICACIÓN MORFOSINTÁCTICA PARA CORPUS DE REFERENCIA EN LENGUA ESPAÑOLA[1]

AURORA MARTÍN DE SANTA OLALLA
Universidad Alfonso X El Sabio
JOSÉ MIGUEL GOÑI MENOYO
Universidad Politécnica de Madrid

0. INTRODUCCIÓN

ESTE TRABAJO presenta una propuesta de codificación morfosintáctica para corpus de referencia en lengua española basada en los estándares de la *Text Encoding Initiative* (TEI), *The Network of European Reference Corpora* (NERC) y *The Expert Advisory Group on Language Engineering Standards* (EAGLES) tal y como se presenta en (Martín de Santa Olalla, 1994). Presentamos también el trabajo de creación de etiquetador morfosintáctico que utiliza el conjunto de etiquetas que ésta contiene.

1. UNA PROPUESTA DE CODIFICACIÓN MORFOSINTÁCTICA PARA CORPUS DE REFERENCIA EN LENGUA ESPAÑOLA

Nuestra propuesta de codificación morfosintáctica para corpus de referencia en lengua española consiste en la creación de un sistema taxonómico que toma como unidad de análisis la palabra ('conjunto de signos entre dos blancos') y describe todos aquellos rasgos que presentan una marca formal explícita que supone, además, un comportamiento gramatical específico.

Nuestro trabajo contiene además un 'manual del codificador' que caracteriza y describe cada una de las clases ('categorías gramaticales' o 'partes de la oración', en nuestro caso).

El punto de partida lo constituyen las propuestas de estandarización para las lenguas europeas: **TEI** (Langedoen y otros, 1991), (Simons, 1991), (TEI-AI-1W2, 1991) y (TEIP3, 1994); **NERC** (Monachini y otros, 1992) y **EAGLES** (Leech y otros, 1994).

TEI se ocupa tanto del contenido como de la forma en el intento de de estandarización de un codificación morfosintáctica.

NERC sólo ha tenido en cuenta hasta el momento la definición de contenidos.

El documento EAGLES, junto al tratamiento de las formas y los contenidos morfosintácticos propone distintos niveles de estandarización en los que se incluyen, aparte de los rasgos exclusivamente morfosintácticos (con sus respectivos valores), ciertos rasgos opcionales de carácter léxico o léxico-semántico separados en dos grupos: aquéllos que son específicos de ciertos trabajos o aplicaciones y aquéllos que son específi-

1. Este trabajo ha sido parcialmente financiado por el Plan Nacional de I+D, por medio del Proyecto: *Arquitectura para Interfaces en Lenguaje Natural con Modelado de Usuario (TIC-910217C02-01).*

cos de ciertas lenguas. Lo más original de este trabajo, en lo concerniente a estandarización, está representado por la propuesta de un nivel intermedio de codificación basado en códigos numéricos en el que los distintos rasgos tienen posiciones fijas en una matriz que es distinta para cada una de las categorías. Esta codificación tiene carácter interlingüístico y constituye un paso intermedio para la codificación de los corpus a partir de la información registrada en un lexicón

Las tres propuestas coinciden en la formalización del análisis morfosintáctico mediante un sistema de pares atributo-valor que se representa por medio de etiquetas o membretes. Las etiquetas o membretes tienen una estructura atómica o jerarquizada que, junto a la pertenencia de las palabras a clases ('categorías') y subclases ('tipos'), refleja rasgos recurrentes y específicos de las distintas formas.

Como resultado de la aplicación de estos estándares a la descripción morfosintáctica de nuestra lengua, nuestro conjunto de etiquetas consta de 660 etiquetas morfosintácticas o 'entidades de segundo orden' que son el resultado de la combinatoria de 117 pares atributo-valor o 'entidades de primer orden'.

Llamamos 'entidades de primer orden' a los pares atributo-valor morfosintácticos. Ellas constituyen la base de nuestra propuesta. 'Entidades de segundo orden' son las estructuras de rasgos formadas a partir de todas las posibilidades de combinatoria de entidades de primer orden que ofrecen las unidades textuales en corpus de referencia en lengua española. Por ejemplo, **C-N** (categoría nombre); **E-4** (propio-sí); **E-5** (propio-no); **G-M** (género-masculino); **G-F** (género-femenino); **N-S** (número-singular); **N-P** (número-plural) son todas ellas 'entidades de primer orden' para la categoría nombre. **N4MS, N4FS, N4MP, N4FP, N5MS, N5FS, N5FP, N5MP** y **N5FP** son todas 'entidades de segundo orden' que reflejan el análisis morfosintáctico de algunas de las formas que puede adoptar la categoría nombre en español.

Las principales clases coinciden básicamente con las 'partes de la oración' tradicionales a las que se añaden las clases 'residual', 'única' y 'puntuación'.

Asignamos la categoría 'residual' a todas aquellas unidades textuales que quedan fuera de las categorías gramaticales tradicionalemente aceptadas y de lo que se consideraría un léxico del español. Su aparición en los corpus es relativamente frecuente y deben ser codificadas. Por ejemplo: palabras extranjeras, fórmulas matemáticas,...

Utilizamos la categoría 'única' para dar cuenta de la codificación de una palabra (o un conjunto reducido de palabras) con un comportamiento específico que la o las hace difícilmente adscribible a algunas de las categorías restantes. Por ejemplo, utilizamos la categoría 'única' para codificar el comportamiento del *que* comparativo en español.

Codificaremos con la etiqueta 'puntuación' todos aquellos signos gráficos que indican límites entre los distintos constituyentes tanto en el marco de la oración simple como en el de la compuesta o en el discurso y sirven además para transcribir distintas entonaciones de un enunciado.

Los atributos codificados en cada clase son los rasgos morfosintácticos específicos de cada clase o categoría gramatical:

- 'propio', 'género' y 'número' para el nombre.
- 'persona', 'género', 'número', 'caso' y 'reflexivo' para el pronombre personal.
- 'forma verbal', 'modo', 'tiempo', 'voz' y 'auxiliar' para el verbo.

Los valores son los correspondientes a los distintos rasgos morfosintácticos ('masculino', 'femenino' y 'neutro' para el género; 'singular' y 'plural' para el número; 'primera', 'segunda' y 'tercera' para la persona,...) más dos valores para subespecificación ('invariante o cualquiera' y 'no-aplicable') y dos valores booleanos Y y O.

Utilizamos 'invariante o cualquiera':

- cuando el rasgo no está formalmente marcado y puede tomar cualquiera de los valores de la escala del rasgo.
- Es un rasgo, sin embargo, que determina un comportamiento morfosintáctico específico.
- No es posible la desambiguación mediante contexto.

Por ejemplo, éste es el valor para el rasgo 'número' en palabras como *crisis*, *chasis* y *martes*.

'No-aplicable':

- el rasgo no es relevante para la palabra que se codifica pero sí lo es en la clase o subclase a la que ésta pertenece. 'No relevante' significa que es un rasgo no marcado formalmente que (y esto es lo que lo diferencia de 'invariante') no determina un comportamiento morfosintáctico peculiar.
- La codificación de este rasgo en una palabra supondría su adscripción a una clase distinta. Por ejemplo, en los adverbios sin significado léxico (los llamados pronominales) la codificación de un valor 'no aplicable' para el rasgo grado, se podría interpretar como la necesidad de establecer una subclase distinta en la que el rasgo grado no fuera un rasgo pertinente.

Por ejemplo, éste es el valor para la concordancia en los participios de las formas compuestas de los verbos o, como acabamos de ver, el del rasgo grado para los adverbios pronominales (*tal vez, quizás, ayer,...*).

O:

- alternancia en una única forma de un subconjunto valores de entre los definidos para ese rasgo. Obsérvese que la diferencia respecto a lo que llamamos 'invariante o cualquiera' es que, en este último caso, la alternancia se daba entre todos los valores posibles para un rasgo. En el caso de los operadores booleanos la alternancia es únicamente entre un subconjunto de los posibles.

Por ejemplo, M|F es el valor del rasgo género en los pronombres personales *yo, me, se,..*

Y:

- conjunción en una única forma de un subconjunto de valores de entre los definidos para un rasgo. Utilizamos esta doble posibilidad de asignación de valores para dar cuenta de la conjunción de un determinado valor formal con otro asignable desde el punto de vista funcional.

Por ejemplo, es la codificación de los rasgos de género y caso en los fenómenos de leísmo, laísmo o loísmo.

2. EL ETIQUETADOR MORFOSINTÁCTICO

El etiquetador que se ha desarrollado pretende ser una herramienta de ayuda al etiquetado morfosintáctico de textos de acuerdo con el esquema de etiquetado que se propone.

El etiquetador está todavía en fase de desarrollo. Su implementación utiliza una metodología de 'prototipado incremental', con la que hasta el momento se han desarrollado dos versiones del etiquetador -la segunda con funcionalidades añadidas a las del primer prototipo. La tercera versión estará disponible en breve. Posteriormente se comentarán las funcionalidades de cada una.

Una de las decisiones de diseño que se tomó fue que el etiquetador entraría en el paradigma de «Sistemas Basados en Conocimiento», principalmente, en conocimiento de tipo lingüístico, sin perjuicio de que en el futuro se puedan integrar diferentes tipos de conocimiento (estadístico, heurístico, etc.).

Para facilitar esta tarea se concibió una arquitectura abierta y modular, en la que cada módulo aporta un tipo diferente de conocimiento. Dichos módulos se invocan cuando es necesario. Por ejemplo, siempre que dos etiquetas son aplicables, o cuando ninguna lo es para una palabra del texto de entrada.

Cuando no hay ninguna evidencia disponible para la asignación de etiquetas, puede invocarse un módulo especial de interfaz con el usuario para que éste, si lo desea, tome la última decisión.

El primer nivel de conocimiento integrado fue el conocimiento léxico. El primer prototipo era capaz de etiquetar el inventario completo de categorías cerradas. Dicha limitación se debía al carácter flexivo de la morfología del castellano que hacía muy costosa la tarea de recoger en el lexicón todas las formas flexivas de los paradigmas verbales y nominales.

La segunda versión del etiquetador incluye un lexicón completo de raíces y morfemas flexivos y un módulo con conocimiento morfológico sobre los fenómenos flexivos en verbos, nombres y adjetivos. Hemos adoptado el modelo y el procesador morfológicos descritos en (Moreno, 1992) y en (Moreno y otros, 1994). En esta versión del etiquetador es necesario consultar frecuentemente al usuario para desambiguar algunas etiquetas puesto que todavía no se tiene en cuenta el contexto sintáctico.

En este momento trabajamos en la integración de un módulo con conocimiento sintáctico que permitiría desambiguar algunas etiquetas. Por ejemplo, si un rasgo no tiene una marca morfológica explícita, su asignación se realiza mediante un fenómeno sintáctico como es la concordancia. De este modo la intervención del usuario podría reducirse enormemente.

La versión actual emplea la plataforma léxica para el castellano descrita en (Goñi y otros, 1994). El lexicón formaliza los rasgos morfosintácticos como pares atributo-valor. Se establece una correspondencia entre las entidades SGML de primer orden que son parte de nuestra propuesta y los símbolos usados para codificar dichos rasgos, que se asocian a una palabra concreta tras un análisis morfológico o directamente desde el léxico.

Esta correspondencia se realiza mediante un módulo especial del etiquetador que puede procesar las especificaciones proporcionadas por el usuario. Por ejemplo, para los nombres, los rasgos de interés son 'género','número' y 'propio' (el resto quedan descartados). La codificación e estos rasgos en el lexicón se muestra en la siguiente declaración:

```
cat = n {agr gen
         agr num
         ninfo proper}
```

Ésta indica al etiquetador que en las entradas en las que el rasgo 'cat' tenga el valor 'n' ('nombre'), los rasgos de interés son 'agr gen', 'agr num' y 'ninfo proper'. Los valores de dichos atributos permiten al etiquetador seleccionar las entidades SGML de primer orden necesarias (columna de la derecha):

cat	= n	C-N
agr num	= sing	N-S
agr num	= plu	N-P
agr gen	= masc	G-M
agr gen	= fem	G-F
ninfo proper	= +	E-4
ninfo proper	= -	E-5

Con estas entidades seleccionadas, el etiquetador puede escoger como candidatas las entidades de segundo orden que incluyan la totalidad de las de primer orden previamente obtenidas.

En este caso son las siguientes:

```
<! ENTITY N5MS "<f.struct> &C-N; &E-5; &G-M; &N-S; </f.struct>">
<! ENTITY N5MP "<f.struct> &C-N; &E-5; &G-M; &N-P; </f.struct>">
<! ENTITY N5FS "<f.struct> &C-N; &E-5; &G-F; &N-S; </f.struct>">
<! ENTITY N5FP "<f.struct> &C-N; &E-5; &G-F; &N-P; </f.struct>">
```

De esta forma, ante un nombre plural y no propio, el etiquetador, una vez seleccionadas las entidades de primer orden **C-N**, **E-5** y **N-P**, propone **N5MP** y **N5FP** como etiquetas candidatas. A partir de este momento la desambiguación debe hacerse mediante otros módulos que integren otro tipo de conocimiento (sintáctico, estadístico u otro) o bien mediante consulta al usuario.

3. CONCLUSIONES

Este artículo presenta una propuesta de etiquetado morfosintáctico para corpus de referencia en lengua española basada en los estándares europeos: TEI, NERC y EAGLES.

Como resultado de este trabajo se presentan un total de 660 entidades de segundo

orden o etiquetas morfosintácticas que son el resultado de la combinación de 117 entidades de pares atributo-valor o entidades de primer orden.

Finalmente, se ha desarrollado un etiquetador morfosintáctico de carácter modular como herramienta informática de ayuda a la asignación de etiquetas. En su versión actual esta herramienta integra conocimiento léxico y morfológico. En este momento se trabaja en la integración de un módulo sintáctico.

BIBLIOGRAFÍA

GOÑI, J. M. y otros (1994): *A Framework for Lexical Representation*. Technical Report UPM/DIT/LIA 5/94. Universidad Politécnica de Madrid.

LANGEDOEN, T. y otros (1991): *Feature-Structure Markup for Presentation at Oxford and Brown Workshops*. Document number TEI AI W9.

LEECH, G.y otros (1994): *MSAL 21 Drafts sections 4.6. and 4.7. of the EAGLES Interim Report: Annotation Subgroup.*

Martín De Santa Olalla, A. (1994): *Una propuesta de codificación morfosintáctica para corpus de referencia en lengua española.* Tesis doctoral. Universidad Autónoma de Madrid.

MONACHINI, M. y A. Ostling: *Towards a minimal standard for morphosyntactic annotation. Technical Report, ILC Pisa. NERC-WP8.2.*

MORENO, A. (1992): *Un modelo computacional basado en la unificación para el análisis y la generación de la morfología del español.* Tesis doctoral. Universidad Autónoma de Madrid.

MORENO, A. y otros (1994): *A morphological model and Processor for Spanish. Technical Report UPM/DIT/LIA 4/94.* Universidad Politécnica de Madrid.

Simons, G. y F. Gary (1991): *Feature System Declarations and the interpretation of feature structures. Document number* TEI AI 1W3.

TEI-AI-1W2 (1991): *List of common morphological features for inclusion in TEI starter set of grammatical annotation tags. Document number TEI AI 1W2.*

TEIP3 (1994): Sperberg-McQueen, C. M. y L. Burnard, editors: *Guidelines for Electronic Text Encoding and Interchange:* volumes I, II. Text Encoding Initiative.